

## Quality and Coverage of Data Sources

### Objectives

Selecting an appropriate source for each item of information to be stored in the GIS database is very important for GIS Data Capture. Selection of quality and coverage of source data is critical because a particular geographic feature is shown on multiple source of varying quality.

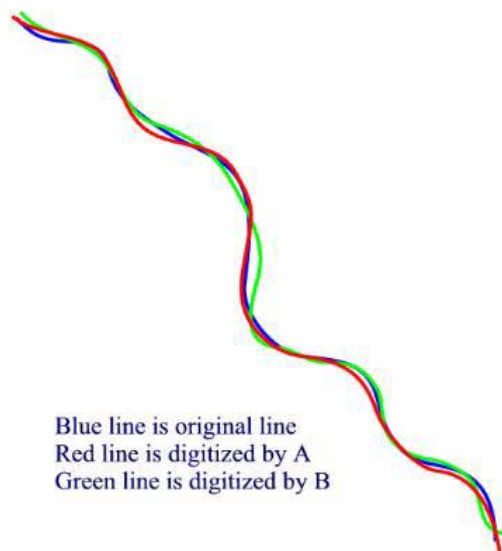
This unit will illustrate and explain how to assess the quality and coverage of source material.

### Key issues involving data sources

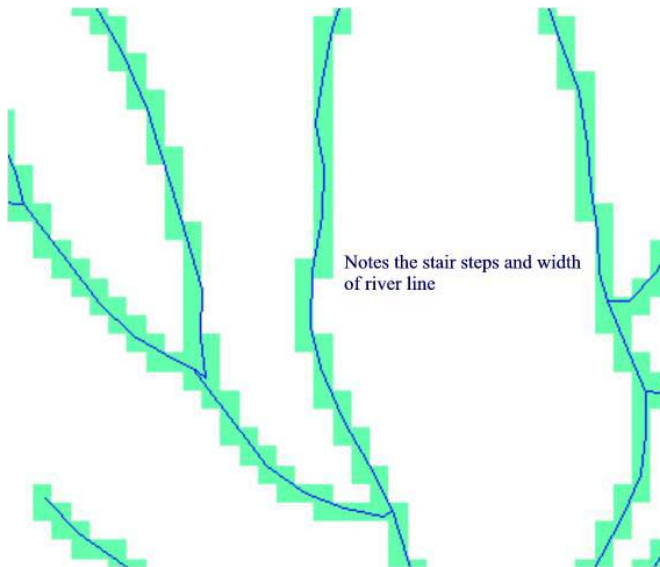
#### 1. Positional Accuracy

The positional accuracy of source map or drawing must be determined or verified. During the verification and determination process, not only the accuracy of source data but also additional error introduced during the digitizing and data input process, must be considered in order to determine the positional accuracy level and establish the quality control error thresholds.

##### I. Positional Error due to Digitizing

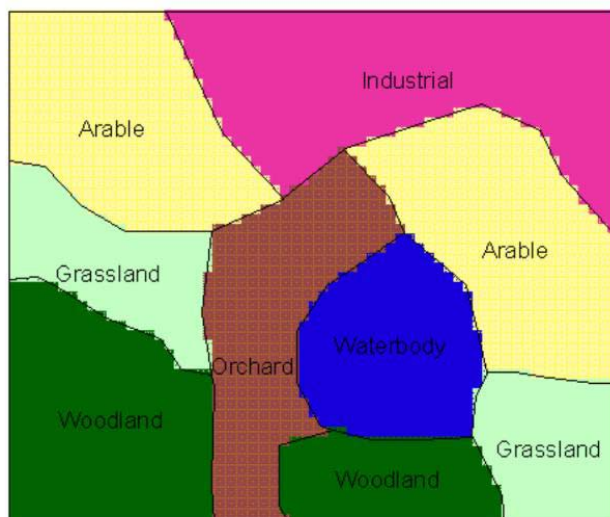


## II. Positional Error due to Raster Conversion



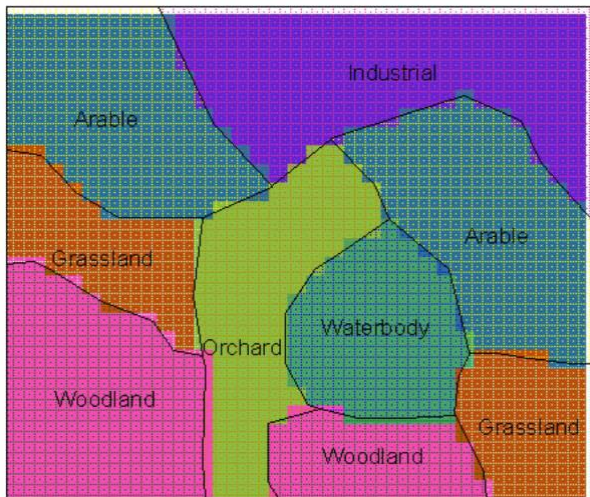
## III. Positional Error due to Raster conversion of Polygons

The detail of polygon boundary is lost.



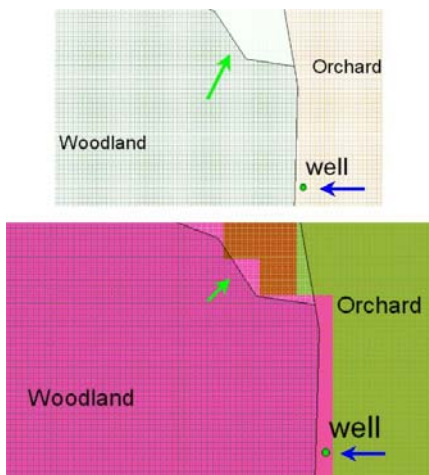
#### IV. Too large grid cell

The mixed pixel problem occurs when grid cells are too large to resolve spatial details. In the following figures, the 100-meter square grid cells are too large to resolve spatial details.



#### V. Inside or Outside

The well is outside the woodland. After conversion to raster, the well is inside the woodland. It is especially important when combining raster and vector databases. See the blue arrow. Moreover, polygon boundary errors are significant at the green arrow.



Generally, the data capturing and input cost increase proportionally with required accuracy. If the data source does not meet the accuracy requirement of application, it may be necessary to find better resources or change the applications.

Comparison to a source of higher accuracy is a recommended practical test of positional accuracy. This approach involves comparison of well-defined points using a set of rules such as the Spatial Accuracy Standard of the American Society for Photogrammetry and Remote Sensing (ASPRS).

The horizontal accuracy and vertical accuracy have to be considered depending on the requirement of the application.

According to the United States National Map Accuracy Standard, Horizontal Accuracy standard is as follow:

For maps on publication scales larger than 1:20,000, not more than 10% of the points tested shall be in horizontal error by more than 1/30 inch or 0.0847 cm, measured on the publication scales.

For maps on publication scales smaller than or equal to 1:20,000, not more than 10% of the points tested shall be in horizontal error by more than 1/50 inch or 0.0508 cm, measured on the publication scales.

The following animation on calculation of acceptable error and allowable RMS implemented the Map Accuracy Standard based on scale.

**See the Animation on the web page.**

These limits of accuracy shall apply in all cases to positions of well-defined points, such as monuments, benchmark, intersections of roads, railroads etc., which can be plotted on the scale of the map within 1/100 inch or 0.0254 cm.

The positional accuracy of different scales can be calculated base on the United States Map accuracy Standard as follow.

1:1,200 ± 100.584 cm  
1:2,400 ± 203.3016 cm  
1:4,800 ± 406.2984 cm  
1:10,000 ± 846.7344 cm  
1:12,000 ± 1015.8984 cm  
1:24,000 ± 1219.2 cm

1:63,360 ± 3218.688 cm  
1:100,000 ± 5080.1016 cm

According to the United States National Map Accuracy Standard, Vertical Accuracy standard is as follow.

Contour maps on all publication scales shall be such that not more than 10% of the elevation tested shall be in error more than one-half the contour interval.

The Spatial Data and representation model of our world in GIS cannot be treated as the perfect model of the world. However, error is inescapable. Therefore error should be treated as a fundamental dimension of data.

## 2. Attribute Accuracy

Categories or polygon feature attributes such as land use, soil and vegetation are representation and adoption of sharp set theory for much more complex environment. These land uses; soil and vegetation etc. data cannot be taken by continuous measure. The categorical polygons or categorical coverage are constructed by setting the attributes as control and measuring the locations of the boundaries between classes. In practice, there is a significant error in identifying these categories on the map due to generalization of complex environment.

With categorical attributes, a class is either right or wrong. To test, it involves determining the category or classification from two sources, and ideally one source should be source of higher accuracy. The quantitative detail method will be discussed in GIS data quality section. The result is a square error matrix, cross tabulating the observed category with the true result. Errors along the row are errors of **omission** with respect to the category and errors along the column are errors of **commission**. It is common to summarize this matrix by the percentage correct, the total of the diagonal. Unfortunately, it is not a reliable index of success across projects with differing frequencies in the various categories. Kappa index, a measure that deflates the percentage correct by the amount, which could be expected to fall into the diagonal under an independent rule of joint probability. Still the raw matrix offers the most complete information to assess fitness for use.

Interactively calculate the Kappa using the following animation.

**See the animation on Kappa calculation.**

Closer the kappa value to 1, the higher the accuracy is.

Moreover, the misclassification matrix is typically obtained through a process such as systematic, random or other spatial sampling scheme for point sampling to generate the sample for testing of accuracy assessment.

Due to the scales and less homogeneity nature of land use, a selected point may not happen to fall purely to a particular land use especially when it involves an area and it can become confused when it is near an edge.

The quantitative detail of misclassification matrix method will be discussed in GIS data quality section.

Alternatively, overlaying a complete second coverage, which is similar scale and resolution or more refined set of categories and more detailed scale to the coverage map to be tested, can generate misclassification matrix.

The standard attribute accuracy in mapping land attributes for land Use and Land Cover Mapping by USGS is as follow.

- 85 percent is the minimum level of accuracy in identifying land use and land cover categories.
- The several categories shown should have about the same accuracy.
- Accuracy should be maintained between interpreters and times of sensing.

### **3. Coverage**

The geographic coverage and scale of source document to the required mapping extent of the GIS project should be inventoried and compared.

Scales and completeness of coverage of project area is often highly variable and patchy. Although, global scale map of some environmental themes such as soil and geology is available, the extent is much more irregular at larger scales.

Normally, other data sources, including field inventory, must be added to complete the intended coverage. Using multiple data sources with different accuracy to complete the database coverage, it is important to alert and trade off to what accuracy level can be reasonably expected within various extent portions of the database.

#### 4. Completeness

Primary source Data should present most, if not all, specific graphic and non-graphic feature data usually in creating a base layer of named objects such as parcels, roads, and census tracts.

Completeness is an evaluation of the existence of all necessary graphic and non-graphic data in the database. For example, roads should not be missing and all roads shown should have a name attached.

The GIS database must be completed with the following components depending on the specific requirement of the project.

- Geometry of spatial features
- Spatial and non-spatial attributes
- Annotation of features
- Connectivity rules among features
- Typology of each feature
- Topological relationship rules for spatial features - Spatial relationship rules for spatial features
- General relationship rules among spatial data and no-spatial data
- Domains (sets of valid attributes values of attributes) and
- Validation rules (that enforce the data integrity through relationship rules and connectivity rules) All facilities depicted on the data source must be captured. For example, a map that is supposed to have 200 parcels is only 75% complete if it shows only 150 parcels. Moreover, the aforementioned components should be completed. Normally, more than one percent of geometry and attributes should not be missed.

Complete data item or attribute must be defined depending on the project requirement. For example, the speed limit, one way, two way, width of the road, number of lanes and turn tables etc. are required data item for network traffic flow analyses.

Secondary and Tertiary sources should be provided to fill the missing or illegible data item from primary source. Secondary and Tertiary source may be field inventory when other acceptable data source do not exist. Inclusion of secondary and tertiary data sources may increase cost and time due to additional document handling and resolving the conflicts of mismatch among the data sources.

Completeness is affected by rules of selection, generalization and scale.

## 5. Timeliness

The spatial features and attribute features change with the time.

It is currentness of the source documents used to create the GIS database. It is important to maintain the currentness of data in the GIS and it reflects the correctness of database because the real world changes daily.

The maps are static representation of spatial features and their respective attributes, as they existed at the time of surveying. The static maps become increasingly inaccurate overtime as the real world of **spatial features change** continuously. Timeliness represents how current the information contained on the source map or document is.

Although, all national mapping agencies maintain a revision or update program, spatial features continue to change. The currency of information is vital for the success of GIS project. Although it is impossible to find a truly current map of an area, every effort should be made to acquire the most up to date information. It is important to note that all features in the map may not be updated at the same time and there may be multiple update cycles for a particular source item.

**Attribute properties of spatial features** are also time dependent. However, updating capability of GIS can update regularly and frequently, which is not possible for a hard copy.

## 6. Correctness

Correctness of valid attributes and valid range of attribute data of GIS database is important.

The location of a spatial object might be correct while its attribute is presented wrongly. A fire hydrant is attributed as parking meter although its location is true.

Subtypes as the spatial attribute can be assigned distinct simple behaviour for



different classifications of features together with Default Values, Attribute Domains, Range Domains, Coded Value Domain, Connectivity Rules and Relationship Rules, which are important aspects of for the correctness of GIS database, database relationship and database intelligence. These will be discussed more detail and integrated way in database management module and intermediate level.

Normally, more than one percent of geometry and attributes should not be incorrectly represented. The geometry and attribute of features should show the correct information that matches the real world geographic feature. A highway should not be shown as river and vice versa. It may be necessary to use field inventory methods to make sure that the data sources are correct. It is very important especially for the network databases such as utility electricity network or fibre optics network. Correctness is essential for GIS data capture and conversion.

## **7. Validity or Logical Consistency**

Validity is a factor to assess whether the source item contains only valid value. It is a measure of valid or permissible values or ranges or conditions. The validity indicates on how well the information depict the source actually reflect the real world conditions.

Internal consistency of the data structure, particularly applies to topological consistency such as unclosed polygons, no label or many labels for each polygons and node-arc errors would be checked.

The validation rules for features can be set up in modern GIS For example: The pH value 20 is invalid attribute value. Topological model can detect and validate the geometric and attribute errors such as missing boundaries, unclosed polygons, over shooting and under shooting at the intersections, too small polygons, too close line and unlabelled polygons. Moreover, the overlay function of GIS is useful to check and avoid placing an electric pole in the river.

Moreover, logical consistency and validity is important for network database connectivity such as water meter can be connected only to water lateral instead of water main. Connectivity rules can be set up to validate the logical consistency of data source and database. Once an error is detected by a validity check, it should be resolved by process of compilation.

## **8. Reliability**

The maps depict significantly different field conditions continually can be considered as unreliable source of information. The cause of the unreliability may be a variety of reasons. However, the map users simply confirmed that the information contains on such maps cannot be reliable based on their experience over time.

## **9. Relevance**

The desired data regarding a site or area to study the particular phenomenon may not exist and surrogate data may have to be used.

Surrogate data, such as Remote Sensing Satellite Image or aerial photograph, is used to measure the phenomenon such as density of particular tree species, or estimate the vegetation cover. The surrogate data is being obtained by an indirect method. Satellite sensors do not see the trees. Sensors record the reflection or emission or back scattering of certain digital signatures typical of trees and vegetation. Sometimes these signatures are recorded by satellites even when trees and vegetation are not present (false positives) or not recorded when trees and vegetation are present (false negatives). Due to indirect measurements of phenomenon, vegetation cover may be over classified or under classified.

It would understand that variations may occur through substituting the surrogate data and although assumptions may be valid, they may not necessarily be accurate.

## **10. Accessibility**

It is important to locate and use the data source easily. Logistical complications often arise. Accessibility to data is not equal. What is open and readily available may be restricted, classified, or unobtainable in another based on the country, scale and content of the map.

## **11. Readability**

It is one of the most important factors of data source quality. The items specified, as originating on a particular source must be readily and consistently legible or readable on that source type.

## **12. Precedence**

A map always shows some information copied from other maps. For example, the land use map showed the roads by copying from the topographic map.

The original information quality will be certainly impacted somewhat in quality. It is preferable to utilize the preceding sources unless second or third generation maps have been maintained the information better than the original map.

## **13. Lineage**

It is a record that addresses the source material or data source from which the data are derived and method or operation of derivation to create the database. It will include method of digitizing, the date and agency of data collection, steps of processing the data and precision of computational results. It is often a useful indicator of accuracy.

## **14. Map Scale**

Each map is produced at a specific scale. Scale is the ratio of units of measurement on the map to units of measurements on the earth. 1: 20,000 scale maps stated 1 cm on the map is equivalent to 200 m on the ground or one unit of measurement on the map is 20000 identical units of measurement on the earth.

1:2000 scale map is larger scale than 1:20000 scale map. A large-scale map covers a relatively smaller area and depicts more detail information than smaller scale map.

As the scale gets smaller, it becomes increasingly difficult to identify the positional errors and any notable errors will be relatively large.

Scale is also important in displaying the map. A GIS can display and plot maps at desired scale. Minimum and maximum scale of display can be set in order to be aesthetically pleasant and cartographically readable.

Scale of the map indicates the level of detail information content and appropriate use of it. 1:20000 to 1:126,720 scale map are used to support land-based mapping to support functions such as general land use planning, airport influence area determinations, environmental management and forestry applications. 1:240 to 1:4800 scale detailed facility maps are appropriate when the smallest potential geographic area of interest is an individual lot or residence.

### Practical Session

1. Fill the appropriate value on the empty raster grid based on the land use value **using the animation on the web page**. Write down your rule on why you fill a particular value to a particular cell. Write down your suggestion to improve the quality in order to maintain the values of original map as much as possible.

2. Calculate the Kappa value based on the following table.

	A	B	C	D	E	F
A	8	0	1	0	2	0
B	0	10	0	2	0	1
C	3	0	21	0	4	0
D	0	0	0	9	0	3
E	4	0	5	0	23	0
F	0	2	0	3	0	14

A = Mixed Forest

B = Paddy

C = Plantation

D = Corn

E = Pine

F = Sugarcane

### Question

1. What are the key issues to be considered to use multi-source data for a project?

### **Share your result**

1. Submit the result of Kappa calculation to the instructor by email.
2. Submit your rule on filling the raster cells by email to the instructor.
3. Submit your suggestion in order to maintain the values of original map as much as possible in filling the raster cell animation exercise.
4. Submit your answer on key issues to be considered to use multi-source data by email to the instructor.